因疫情影响，大作业环境采用线上模型，Atlas 200 DK环境目前仅部署一台线上机器。由于我们班的同学人数较多，为加快同学们的实验进度，现提供本地CPU设备推理的方案。

## 华为云实验

在完成【ModelArts训练修改版.docx】文件的内容后，在MindSpore的版本是1.5.2的条件下，可以导出得到mobilenetv2.onnx文件。经查询，该ONNX模型的opset version是1.11.0。

可使用Netron工具可视化ONNX模型。

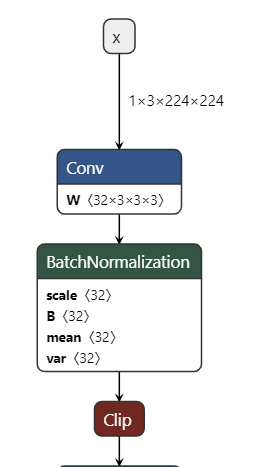


图 1 ONNX模型部分可视化结果

## 本地推理

### 环境配置

在实验3 MindSpore环境的基础上，安装onnx和onnxruntime包，命令如下：

pip install onnx==1.10.0

pip install onnxruntime==1.10.0

这里使用1.10.0版本的原因是，Clip算子在opset version为11时，min、max属性被移除，而是作为inputs，可访问<https://github.com/onnx/onnx/blob/main/docs/Changelog.md#Clip-11>查看。但是，导出的ONNX文件的Clip算子是将min、max作为attribute，所以opset version的值应该<11。

### 模型检查和修改

使用checker检查导出ONNX模型会发现Clip算子在opset version为11的情况下出现错误，所以修改model的opset version。

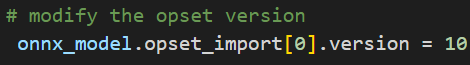


图 2 修改ONNX模型的opset version

直接进行推理，会提示出现GEMM算子输入的rank不等于2，经检查，发现ReduceMean算子的keepdims属性默认设置为1，即计算均值后的结果还是保持与输入一致的维度，这是引起了GEMM算子的输入维度错误。因此，需要修改keepdims属性的值为0，令ReduceMean的输出rank为2。具体代码如下图所示，流程是寻找ReduceMean算子，替换该算子。

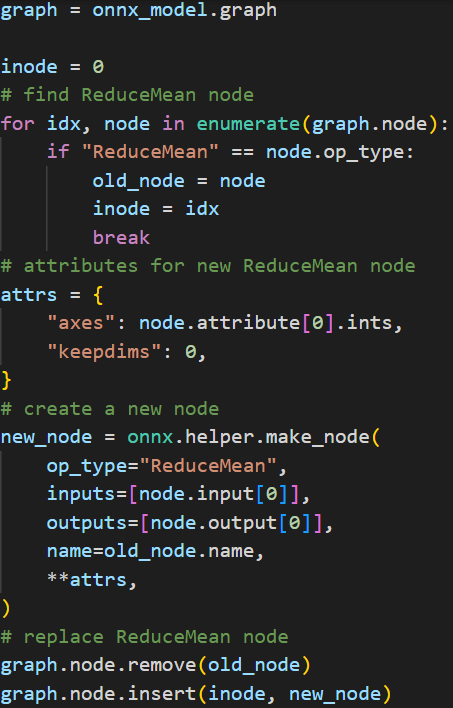


图 3 修改ReduceMean算子

### 图像预处理

因为使用Atlas 200DK时，引用了atlas\_utils文件夹的代码用于对图片做预处理。而在丢弃Atlas 200DK环境时，我们需要更改图片预处理的代码。图片预处理的作用是：修改输入图片的维度以匹配模型的输入维度、令输入矩阵的排列格式是CHW以匹配模型输入。所以基于PIL、numpy库，以下面的代码完成这两个功能。

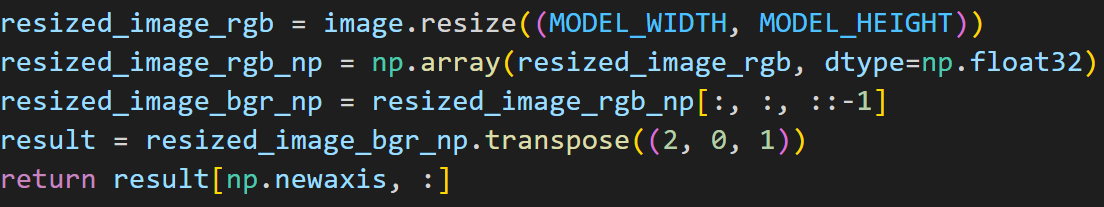


图 4 图片预处理的代码

其中，np.newaxis是为了增加一个维度，即令排列格式为NCHW，其中N是batch\_size。

### 模型推理

基于ONNX格式的模型推理步骤较为简单。在读取ONNX模型和经过checker检验后，创建一个推理的session，然后读取图片，进行预处理，调用run接口进行推理，然后使用后处理代码把预测结果draw到原始图片中。

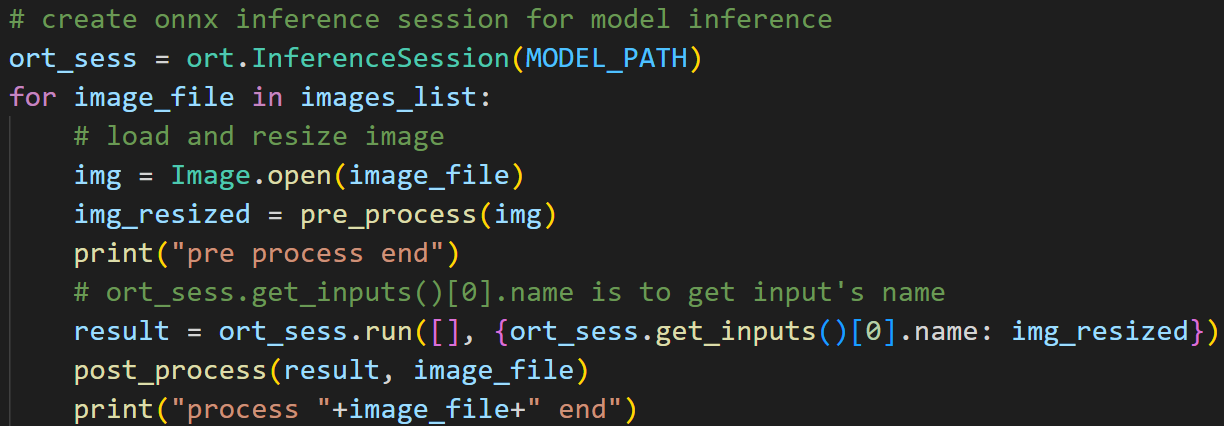


图 5 ONNX模型的推理代码

其中post\_process的代码与使用Atlas 200DK编写的classify\_test.py一致。推理结果如下所示：



图 6 推理结果示例

## 后续扩展

ONNX是一种针对机器学习所设计的开放式的文件格式，它可用于多种边缘设备上进行推理。可以考虑对ONNX模型进行量化、模型压缩等加速操作，对比前后的耗时变化；也可以对比在不同设备上ONNX模型推理速度的变化等。

实验代码已开源在<https://github.com/panlinchao/onnx-inference>。